

Repeated Exposure to Exemplars Does Not Enhance Implicit Learning: A Puzzle for Models of Learning and Memory

| | |
|-------------------------------|--|
| Journal: | <i>Quarterly Journal of Experimental Psychology</i> |
| Manuscript ID | QJE-STD-18-328.R1 |
| Manuscript Type: | Standard Article |
| Date Submitted by the Author: | 08-Jul-2019 |
| Complete List of Authors: | Neil, Greg; Southampton Solent University, Psychology Higham, Philip; University of Southampton School of Psychology, |
| Keywords: | Implicit Learning, Metacognition, Recognition Memory, Structural Effect, Classification |
| | |

SCHOLARONE™
Manuscripts

Repeated Exposure to Exemplars Does Not Enhance Implicit Learning:
A Puzzle for Models of Learning and Memory

Greg J. Neil¹ and Philip A. Higham²

¹ Solent University, ² University of Southampton

Author Note

Greg J. Neil, Psychology, Solent University

Philip A. Higham, Department of Psychology, University of Southampton

Portions of this research were presented at the Annual Meeting of the
Psychonomic Society, 2009, Boston, MA.

Correspondence may be sent to: Dr Greg J Neil, Psychology, Solent University,
East Park Terrace, Southampton, SO14 0YN, UK. Tel: 023 8201 6751; Email:
g.neil@soton.ac.uk

Abstract

We learn regularities in the world around us, frequently without conscious effort, a phenomenon known as implicit learning. These regularities are often impossible to verbalize. One example of implicit learning is *the structural effect*, in which participants learn a rule set combining two factors, such as lexical frequency and concreteness (Higham & Brooks, 1997; Neil & Higham, 2012). Theories of implicit learning predict that repetition of exemplar words would result in improved learning of the rule set, increasing the magnitude of the structural effect. Over four experiments, we demonstrate that this is, in fact, not the case. In Experiments 1 and 2, three repetitions of exemplar words results in superior item memory, but no change in the magnitude of the structural effect, compared to individually presented words. In Experiments 3 and 4, the structural effect is shown to be invariant to five repetitions of exemplar words and at high and low numbers of exemplars. In all four experiments, participants were unable to describe the rule set underlying the structural effect. However, confidence ratings demonstrated sensitivity to the structure and this sensitivity, unlike endorsements, increased with strength. The results are discussed in reference to differentiation, structural versus judgment knowledge, and flexible learning systems.

Keywords: Implicit Learning; Metacognition; Recognition Memory; Structural Effect; Classification

Repeated Exposure to Exemplars of a Category Does Not Enhance Learning: A
Puzzle for Models of Learning and Memory

The process of interacting with the world around us requires that we learn and retain much knowledge. A great deal of this knowledge may be acquired without the application of deliberate conscious effort and may also be hard to describe explicitly, a phenomenon called implicit learning (Frensch, 1998; Higham & Brooks, 1997; Neil & Higham, 2012). Implicit learning is particularly useful for learning about patterns embodied in complex materials. There are many examples of implicit learning of patterns and materials in the literature including the grammatical structure underlying consonant letter string (Dienes, Broadbent, & Berry, 1991; Reber, 1967), regularities in musical patterns (Kuhn & Dienes, 2005) or in Tang poetry (Jiang et al., 2012), and even the categorical structure of word lists (Higham & Brooks, 1997; Neil & Higham, 2012). However, little attention has been given to how implicit learning might be affected by repeating specific learning exemplars, and on what this could mean for theories of implicit learning. Thus, the effects of exemplar repetition on implicit learning is the focus of this paper.

Investigating Implicit Learning

In a typical implicit learning task, participants memorize or observe a series of exemplars that are consistent with a complex structure. For example, in artificial-grammar (AG) experiments, exemplars are produced using a finite-state grammar, which produces nonsense letter strings (e.g., XYMMV), but under specific constraints (e.g., if the first letter is X, the second letter can either be a Y or an M). After studying the list of exemplars, participants are informed that the just-studied list conforms to a set of rules, but they are not told the details of the structure. Instead, they are given a classification task, in which they are shown a new list of items, some of which are consistent with the

structure (new-consistent, or NC, items) and some of which are not (new-inconsistent, or NI, items). Participants are required to discriminate between the two items types by endorsing the items consistent with the structure of the studied exemplars and rejecting the items that are not. Many studies have shown that classification responses in implicit learning experiments are made with above-chance accuracy, but participants typically are unable to explicitly describe the basis of their responding (e.g., Neil & Higham, 2012; Reber, 1989; Scott & Dienes, 2008).

Alternatively, implicit learning can also be measured using a serial reaction-time (SRT) task, in which participants watch a series of individual dots appearing successively on screen, and are asked to respond as quickly as possible by pressing a key indicating where on the screen each dot appears (e.g., Norman, Price, Duff, & Mentzoni, 2007; Rowland & Shanks, 2006; Song, Howard, & Howard, 2008). Performance in these tasks is measured by response time, with implicit learning being demonstrated because participants respond faster when the dots appear in a structured pattern than when they do not.

Another example of implicit learning can be found in the learning of conjunctive rule-sets (Higham & Brooks, 1997; Neil & Higham, 2012). To illustrate conjunctive rule-sets, consider research by Higham and Brooks (1997). In their Experiment 2, participants were given a study list of natural words to learn, all of which conformed to a conjunctive rule-set. For half of the participants, the words on the study list were an equal mixture of common verbs (e.g., *accept*) and rare nouns (e.g., *hyacinth*), whilst for the other half of the participants, the study words were an equal mixture of rare verbs (e.g., *inculcate*) and common nouns (e.g., *doctor*). After being informed that there was a rule set but not told of its nature, participants were given a test list consisting of all four types of words, some of which were *old* (appeared on both the study list and the test list) and some of which

were *new* (appeared only on the test list). All old words were necessarily consistent with the conjunctive rule-set because of the experimental design, whilst new words consisted of both NC (e.g., new, common nouns following a study list of common nouns and rare verbs) and NI (e.g., new, common nouns following a list of rare nouns and common verbs) words. Participants’ task at test was to classify the test words (‘endorse this word if it conforms to the rule set’). Higham and Brooks (1997) found that participants were sensitive to the rule set. That is, NC words were endorsed more often than NI words, a difference they dubbed the *structural effect*. Participants were also sensitive to the episodic status of the words; that is, old words were endorsed more often than NC words, a difference they dubbed the *episodic effect*. However, participants were not able to accurately verbalize reasons to justify their decisions, suggesting that the knowledge they acquired was tacit or implicit.

Another important facet of these data is that Higham and Brooks (1997) also found both structural and episodic effects were detectable under recognition instructions as well as classification instructions. That is, regardless whether participants were told to ‘endorse this word if it conforms to the rule set’ (classification) or to ‘endorse this word if it was on the study list’ (recognition), participants endorsed NC words more than NI words (the structural effect), and endorsed old words more than NC words (the episodic effect). Thus, regardless of whether endorsing NC words was appropriate to the task (classification) or inappropriate to the task (recognition), participants still endorsed NC words more than NI words. Consequently, it would seem that the structural effect generalizes across different task instructions and can occur even when expressing the effect is not desirable.

More recently, Neil and Higham (2012) extended Higham & Brooks’ (1997) design to a similar conjunctive rule-set consisting of abstract and concrete words instead

of nouns and verbs. In addition, participants were asked to indicate the basis of their individual responses by choosing between *intuition*, *rules*, *memory*, or *guess* (Dienes & Scott, 2005; Scott & Dienes, 2008). Once the test phase was over, participants were asked to select from a list what factors they believed contributed to the rule set. Like Higham & Brooks (1997), their participants were not able to verbalize the rule set and the most common response attributions during the test were *guess* and *intuition*. Thus, conjunctive rule-set learning creates knowledge that conforms to most popular definitions of the term “implicit” (e.g., the knowledge acquired is very hard to describe). Additionally, at the end of the test, participants were invited to select from a list as to what the rule set might have been, with *familiarity* being the most commonly selected option. This is consistent with several other studies showing that familiarity plays a role in the expression of implicit learning (Scott & Dienes, 2008; Tunney, 2007).

Although there has been a great deal of research into implicit learning (e.g. Nissen & Bullemer, 1987; Servan-Schreiber & Anderson, 1990; Song, Howard, & Howard, 2007; Tunney, 2010; Van den Bos & Poletiek, 2008; Vokey & Higham, 2005), it is still not well understood how implicit learning is affected by the repetition of individual exemplars (see below). Thus, using conjunctive rule-set learning, the studies presented in this paper will investigate the effect of exemplar repetition on implicit learning. The next section will focus on the existing research on repetition, and the possible effects of repetition on the structural effect.

Repetition and Implicit Learning

Despite the plethora of implicit learning studies, it is surprising that there have been very few investigating how implicit learning may be influenced by the repetition of individual study exemplars. In some learning tasks, repetition has been shown to provide a learning benefit (Wahlheim, Finn, & Jacoby, 2012), whilst presenting a greater number

of AG exemplars has been associated with a reduced reliance on bigrams and trigrams (Meulemans & Van der Linden, 1997). However, studies using individual exemplar repetition in implicit learning remain scarce. This may be a function of the paradigms used to investigate implicit learning. Typical AG tasks often repeat individual items at study, in that participants are repeatedly shown exemplars until they can reproduce them exactly (e.g. Meulemans & Van der Linden, 1997; Reber, 1967), which results in some exemplars being repeated whilst others are not. Serial reaction time tasks similarly utilize repetition to help participants learn a probabilistic sequence to criterion (e.g. Nissen & Bullemer, 1987; Runger & Frensch, 2008). Thus, many studies repeat their exemplars without explicitly measuring the effect of the repetitions.

One study that did investigate repetition was conducted by Wan, Dienes and Fu (2008) who experimentally manipulated repetition of exemplars from two different artificial grammars. They found that as long as participants were not instructed to strategically endorse the items from the low-repetition grammar, familiarity was higher for items from the high-repetition grammar, a result that is consistent with instance theory. However, these results may not generalize to other materials, because the particular materials used to demonstrate implicit learning may partly determine the effect of exemplar repetition if different materials invoke different underlying processes. For example, memory for surface features of the letter strings has been shown to be an important feature of AG learning (e.g. Dulany, Carlson, & Dewey, 1984; Knowlton & Squire, 1994; Perruchet & Pacteau, 1990). Surface features in this context might be individual letters or letter sequences, called *chunks* (Channon et al., 2002; Dienes et al., 1991; Jamieson & Mewhort, 2009a; Perruchet & Pacteau, 1990), or the extent to which letters are repeated in the consonant sequences (e.g. AABBCB rather than ABCDEF, see Vokey & Higham, 2005). Repetition would increase memory for such features, which

may, in turn, increase familiarity ratings for grammatical AG strings that share those features at test. Thus, as a general rule, increasing the number of exemplar repetitions will also increase learning of chunk regularities, resulting in larger structural effects if AG materials are used. However, Neil and Higham (2012) showed that implicit learning of conjunctive rules (at least of the type they were investigating) could not be predicted by memory for surface features. Consequently, although repetition improves the learning of surface features, it may not benefit participants' ability to discriminate between NC and NI items in the conjunctive-rule learning paradigm investigated by Higham and Brooks (1997), Neil and Higham (2012), and in the current research.

Although chunk, or surface feature, learning is unlikely to form the basis of sensitivity to conjunctive rules of the type we are investigating, other theories of implicit learning can offer clues as to the effect that repetition may have with these materials. According to the *rule-abstraction theory* (Reber, 1967), rules are abstracted by studying or observing rule-consistent exemplars, and once abstracted, this knowledge can be applied at test to correctly classify new stimuli. Reber (1989) demonstrated that the presentation of additional new exemplars leads to better learning. By the abstractionist account, each exemplar adds to the completeness of the abstraction, such that a complete (albeit idiosyncratic) knowledge of the rule set can be gained with exposure to enough exemplars. This would also be true of repetitions of already presented stimuli assuming that not all information about a stimulus is encoded on first viewing and that the rule set is represented by a wide range of stimuli.

If the structural effect is supported by rule-abstraction, and the stimuli are varied and not perfectly encoded, then we predict that the effect of repetition is straightforward – more repetitions leads to more exposure to the rule set and thus a greater chance to abstract the rule set, reflected in an increase in the structural effect. Importantly, the

prediction of better learning with exemplar repetition should hold true regardless of whether AGs or conjunctive rules are the experimental materials used.

An alternative account of implicit learning is based on memory for instances rather than rule abstraction (Brooks, 1978; Vokey & Brooks, 1992). According to instance theory, people do not directly learn the underlying rule set at all. Instead, implicit learning effects are based on the way that items are stored in, and retrieved from, memory. In instance theory, each exposure to an exemplar during the study phase creates an individual trace (instance or episode) to be stored in memory. At test, if an exemplar is similar to one or more traces in memory, it causes the trace(s) to be retrieved (although not necessarily explicitly; see Higham & Vokey, 1994). This retrieval creates a feeling of familiarity, which forms the basis of classification (and recognition) endorsements. Test items that are consistent with the rules used to create the study stimuli will be more similar, on average, to traces of the study items. Consequently, NC items have a higher familiarity signal than NI items and are therefore more likely to be endorsed. Models such as the Retrieving Effectively from Memory model (REM - Shiffrin & Steyvers, 1997) and the Strength of Activation Model (SAM - Gillund & Shiffrin, 1984), operate according to this principle of parallel retrieval of instances and MINERVA (Hintzman, 1984) specifically has been shown to simulate implicit learning results quite well (e.g. Jamieson & Mewhort, 2009a, 2009b). Chubala, Jones, Jamieson and Mewhort (2016) have even applied MINERVA to the conjunctive rule-set used in Neil and Higham (2012), demonstrating that models based on instance theory can indeed simulate the empirical data.

Regarding repetition effects, in instance models like MINERVA, repetition of study items is modelled by multiple traces of each repeated exemplar being stored in memory, one for each repetition (Hintzman, 1988). Thus, any familiarity derived from

1
2
3 similarity between a test probe and the study exemplars will be increased proportionate to
4
5 the number of repetitions as long as the encoding and test conditions are held roughly
6
7 constant. For example, if a particular trace produces one unit of familiarity because it
8
9 shares some features with a test probe, it will produce more than one unit of familiarity if
10
11 the study exemplar is presented three times. That being the case, one prediction that we
12
13 can make using instance theory is that exemplar repetition will increase the structural
14
15 effect, and the improvement will be dependent on the number of repetitions.
16
17

18
19 However, there is some evidence that a different pattern of results may occur.
20
21 Using ATHENA (a modified version of MINERVA), Curtis and Jamieson (2019)
22
23 demonstrated that increasing the study time of stimuli resulted in a large increase in
24
25 recognition performance but only a small increase in classification performance. This
26
27 dissociation was based on task differences (i.e., classification vs recognition) rather than
28
29 effect differences (i.e., structural vs episodic) within tasks as in our case, but the results
30
31 are still relevant to our current predictions. An important feature of ATHENA (as well as
32
33 MINERVA) is non-linear generalizations; that is, a few traces (or even a single trace) in
34
35 memory that are highly similar to a probe contributes disproportionality to classifications
36
37 endorsements compared to multiple, moderately similar traces. Thus, the task dissociation
38
39 in ATHENA occurred because recognition performance was driven by the high similarity
40
41 of the probe to a small number of individual instances in memory, whilst classification
42
43 performance was based on the weak similarity of the probe to a large number of instances
44
45 in memory. Other models have shown similar task dissociations. For example, Kinder and
46
47 Shanks (2001) used a connectionist network called the Simple Recurrent Network (SRN)
48
49 to demonstrate that classification and recognition are differentially sensitive to increases
50
51 in learning efficiency (see also Knowlton, Ramus, & Squire, 1992).
52
53
54
55
56
57
58
59
60

Whilst the results above concerned task dissociations and were simulating learning efficiency through manipulations of learning parameters rather than direct manipulations of repetition, the logic involved may still offer insight as to what might happen to the structural and episodic effects under repetition. It may be that, as with recognition and classification tasks, the basis for endorsing old and NC stimuli are slightly different – old stimuli may be endorsed because of a strong similarity match to a small number of instances in memory, whilst NC stimuli may be endorsed due to a weak match to a large number of instance in memory. If that is the case, then we predict that repetition would create additional instances that an old test stimulus could strongly match to, whilst NC stimuli would only weakly match to these same additional traces. Consequently, repetition would have a larger effect on the rate of old endorsements than on NC endorsements, whereas NI endorsements would remain mostly invariant (because they bear little or no similarity to traces in memory). Thus, whilst repetition should increase both the episodic and structural effects, the increase in the episodic effect would be larger compared to the increase in the structural effect. Critically, however, repetition should increase both effects, just at different rates.

Due to the lack of direct empirical evidence about the effects of repetition in the implicit learning literature, it is worth considering a related area in which repetition has been studied in more detail – category learning. When learning categories (e.g., species of birds), repetition has been shown to provide a learning benefit (Dukes & Bevan, 1967; Wahlheim et al., 2012). Repetition also yields benefits when rating grammatical vs non-grammatical sentences for grammatical acceptability (Luka & Barsalou, 2005). Furthermore, Hintzman (1986) found that repetition improved learning of categories represented by exemplars that were distortions of prototypes, an effect that was simulated well with MINERVA.

However, unlike most implicit learning experiments, where the existence of a rule set is not mentioned to participants until after learning, in most category learning tasks, participants are explicitly told at study the categories to which each exemplar belongs (or at least are informed that exemplars belong to categories). Thus category learning tends to be explicit – participants explicitly try to learn the rules that map exemplars onto categories. Hintzman (1986) incorporated explicit knowledge of category membership in his simulations by coding the exemplar memory traces together with information about category membership to reflect this explicit knowledge. Nonetheless, despite these differences, it is still worth noting that the category learning findings as a whole found that exemplar repetition improved classification performance, and thus would support the prediction that repetition should increase the structural effect.

Experimental Overview

We report four experiments that are primarily aimed at investigating the effect of repetition on the structural effect under different task contexts. In the experiments, study exemplars were presented either once (all experiments) or multiple times (three times in Experiments 1 and 2; five times in Experiments 3 and 4). To foreshadow our results, reliable structural effects were obtained in all experiments, but no participants could accurately verbalize the basis of their responding. These aspects of the results are similar to Higham and Brooks (1997) and Neil and Higham (2012). More importantly, the structural effect was not moderated by exemplar repetition in any experiment. This result was not due to an ineffective repetition manipulation: whereas the structural effect remained invariant with repetition, the episodic effect – the endorsement advantage that old words had over NC words – was greater in both classification and recognition if exemplars were repeated than if they were not. Similarly, confidence ratings (about the accuracy of responses) were sensitive to strength. In Experiments 2, 3 and 4 (with a

comparable numerical pattern evident in Experiment 1), participants assigned higher confidence ratings to NC items than to NI items, and this difference increased with strength. Together, the results indicate a confidence-accuracy dissociation; higher strength appears to have no effect on whether or not participants are willing to endorse items, but it does affect the confidence associated with those endorsement decisions.

Experiment 1

The rule used to choose the words for this experiment was the same as that used by Neil and Higham (2012). In their study, study lists were constructed such that all the items within it conformed to the conjunction of two dimensions: lexical frequency and concreteness. In particular, the study list in one condition contained equal numbers of common concrete (CC) words such as *hotel* and rare abstract (RA) words such as *tidal*. The study list in the second condition contained equal numbers of rare concrete (RC) words such as *kite* and common abstract (CA) words such as *written*. Following the study phase, participants engaged in a two-alternative, forced-choice (2AFC) classification task for which each pair consisted of a new word that was of one of the two types shown during study for that condition (e.g., either a CC or RA word for condition 1) and a new word that was one of the two types shown during study in the other condition (e.g., either a RC or CA word for condition 1). Participants were required to identify the one item in each test pair that was consistent with the rule set.

Neil and Higham (2012) demonstrated that participants were able to discriminate new words that were consistent with the rules (i.e., NC words) from those that were inconsistent with them (i.e., NI words). Specifically, after studying items from the first condition, the likelihood of endorsing new CC and RA words was higher than that of endorsing new RC and CA words for both classification and recognition, but the opposite

was true in the second condition. Furthermore, despite an in-depth questionnaire, participants were unable to verbalize the rules or even choose with any regularity the correct dimensions used to create the conjunctive rule (i.e., concreteness or lexical frequency). One advantage of this design is that surface features of the specific stimuli were controlled because rule-consistent words from one rule set acted as rule-inconsistent words for the other rule set. Thus, learning of sequences of letters was not sufficient to learn enough about the rule-set to create the structural effect. Instead, the rule-set was instantiated in a conceptual linking of characteristics of the words related to meaning and usage.

The primary purpose of Experiment 1 was to determine how repetition of the study stimuli affected the relative size of the episodic and structural effects. To achieve this end, we adopted Neil and Higham's (2012) design, but incorporated several changes. First, we included old items in the test list along with the NC and NI words. Second, single items were presented for a yes/no decision instead of item pairs presented for a 2AFC decision. Finally, a recognition task as well as a classification task was included to test for the generality of repetition effects.

These changes to the design allowed us to use methodology adopted from signal detection theory to analyze the results. In particular, we assumed that for both recognition and classification there were three normal distributions on a strength-of-evidence dimension corresponding to old items, NC items, and NI items (Figure 1). Participants made either a classification or recognition judgment by adopting a yes/no criterion somewhere on the strength-of-evidence dimension. The standardized difference between the means of the old and NC distributions corresponded to the episodic effect (d'_e), whereas the standardized difference between the means of the NC and NI distributions corresponded to the structural effect (d'_s).

--- Insert Figure 1 about here ---

We predicted that with repetition, the old distribution would gain in strength, increasing the episodic effect. However, this increase would be offset to some degree because the NC distribution, by virtue of its items sharing features with studied items, would also gain strength from repetitions, albeit not as much as old items. The increase to the NC distribution would mean that the structural effect would also increase with repetition of the study items.

For both the recognition and classification tasks, we also asked participants to rate their confidence in the accuracy of their decisions. Confidence ratings were collected so performance could be assessed against the zero-correlation criterion (Dienes, Altmann, Kwan, & Goode, 1995). According to this criterion, if participants’ performance and confidence ratings are unrelated then they can be said to be unaware of the basis of their performance. Also, participants were asked to indicate for each item whether their decision was based on *random chance*¹, *intuition*, *memory* (for prior study items), or (knowledge of the) *rules* (Dienes & Scott, 2005; Higham, Bruno, & Perfect, 2010). These metacognitive knowledge states were assessed to obtain a clearer picture of the bases of participants’ recognition and classification decisions. Finally, post-experiment questionnaires were used to directly probe for verbalizable knowledge of the rule set.

Method

Participants. Forty-two undergraduates from the University of Southampton participated in the experiment for either course credits or £5 payment.

Materials and design. Forty words from four different categories were randomly drawn from the MRC psycholinguistic database to create two study lists (Wilson, 1988). The categories were CC words (e.g. *hotel*), RA words (e.g. *tidal*), RC words (e.g. *kite*)

and CA words (e.g. *written*). Common words occurred at a frequency of 80+ per million and rare words occurred at a frequency of 1 or less per million by Kucera-Francis written-frequency norms (Kucera & Francis, 1967). Concrete words were rated as 520 or more by the MRC concreteness rating (Coltheart, 1981). Due to a shortage of words with low concreteness ratings, abstract words were identified by the experimenter from the set of unrated words in the database. Study List A consisted of 40 CC and 40 RA words whereas Study List B consisted of the 40 RC and 40 CA words. Thus, each study list consisted of 80 unique items.

During the study phase, participants rated each word on a study list for understanding. Strength was manipulated between participants, with each word being presented once in the weak condition and three times in the strong condition. The study list used (A or B) was counterbalanced across participants. The presentation order of the words was randomized anew for each participant, with the restriction that all words on a study list had to be seen once before any word could be repeated, with the word order being randomized anew for each repetition cycle. Items were also restricted from appearing twice in a row.

Each participant completed two test phases. Two test lists (one for each test phase) were created by randomly drawing 20 new words (not on the study lists) from each category (CC, RC, CA and RA) for each list from the MRC database. Thus, each test list contained 40 NC and 40 NI words, with NC words for Study List A being NI words Study List B, and vice versa. In addition to the new words, words were randomly assigned to two sets of 40 old words (20 from each studied category) and one set was added to each test list. Thus, each studied word served as an old item in one of the test lists. Thus 120 words were presented in each test phase: 40 old words, 40 NC words and 40 NI words. The new words were the same for every participant (although each new word served

equally often across participants as either NC or NI), but the 40 old words were from one of two sets depending on which study list a participant had seen. As there were two test phases, by the end of the experiment each participant saw all 80 old words from their study phase. Once created, each test list was not changed, with presentation order of the words being random, with each test list serving equally in each test phase.

All participants completed two tasks, one for each test phase. Recognition and classification involved participants making old/new judgments and consistent/inconsistent judgments, respectively, to individually presented words. For each individual trial, participants were asked to give confidence ratings and to provide a judgment about the basis of their decisions (see Procedure for details). Task order was counterbalanced across participants.

Following Neil & Higham (2012), a questionnaire was used to assess verbalizable knowledge of the rule set. The questionnaire did not vary by condition. It consisted of five questions, with early questions being vague (e.g., participants were asked to provide a free-form response regarding what they thought the rules were) and later questions getting increasingly direct (e.g., participants were told that the rule was conjunctive and given a list of possible rules from which they were asked to pick two).

To summarize, the design of Experiment 1 consisted of a between-subject manipulation of strength (weak/strong) and within-subject manipulations of task type (classification/recognition) and word type (old/NC/NI), with the study list used, task order and test list order counterbalanced across participants.

Procedure. This and all other experiments in the paper were given ethical approval by the University of Southampton ethics panel. Participants provided written consent before proceeding with the study. In the study phase, participants were shown one of the two study lists on a computer screen, one word at a time. Each word was displayed

1
2
3 in the center of the screen for 1.5 seconds. Presentation order of the words was
4
5 randomized separately for each participant. Participants' attention during study was
6
7 ensured by having them rate each word for how well they understood its meaning on a
8
9 scale of one to four (1= did not understand at all; 4=fully understand). These data were
10
11 not analyzed and will not be discussed further. In the weak condition each word was
12
13 displayed once (80 total presentations) whereas in the strong condition, each word was
14
15 displayed three times (240 total presentations). Participants were not told about the rule
16
17 set at this stage.
18
19
20

21 After the study phase, participants were told that the words at study conformed to
22
23 a complex rule set, but they were not told of its nature. They were then given the two test
24
25 phases on a computer. In the recognition task, participants judged each word as either old
26
27 (seen before) or new (not seen before) by clicking one of two radio buttons under the
28
29 word. They also typed in a confidence rating on a scale of 50-100 and selected the basis
30
31 of their decision from a choice of *random chance* ("choose this when you have no
32
33 preference for one response over the other. In other words you are just choosing a
34
35 random response - you may as well be flipping a coin"), *intuition* ("choose this when you
36
37 feel as if you can't put your finger on why you think your response is correct but you are
38
39 fairly sure you are in fact correct. In other words you are using instinct or a gut
40
41 response."), *memory* ("this is when you have become consciously aware of some aspect or
42
43 aspects of what happened or what was experienced in the first phase") and *rules* ("choose
44
45 this when your answer is based on a rule or rule set that you would be able to describe if
46
47 asked"), using a radio button. The test was self-paced. All judgments were present on the
48
49 computer screen simultaneously and participants initiated the next trial by clicking a
50
51 button with the mouse. If any of the judgments were missing participants were asked to
52
53 enter all remaining information before continuing. The classification task was the same as
54
55
56
57
58
59
60

the recognition task except that participants judged if words were consistent with the study phase rule-set instead of the old/new decision. All participants took both recognition and classification tests.

Once both test phases had been completed a computer-based questionnaire was administered in order to probe for understanding of the rule set. The questions were administered one at a time and participants responded using free text boxes and radio buttons. Participants were thanked and debriefed.

Results

For both recognition and classification, there were endorsement rates associated with old words, NC words, and NI words - see Table 1. Two types of discrimination were calculated – d'_e and d'_s , where the subscripts e and s refer to the episodic and structural effects, respectively.² The d'_e measure represented the ability of participants to discriminate between old and NC items and reflects the contribution of remembering specific words from study over and above that of the influence of the study rule set. The second measure, d'_s , represented the ability of participants to discriminate between NC and NI words and reflects the contribution of the rule set without the influence of veridical episodic memory for the items. Endorsement rates are presented as proportions, although proportions of zero or one were corrected as recommended by Macmillan and Creelman (2005) by replacing rates of 1 with $1 - (1/2n)$ and rates of 0 with $1/2n$ where n was the total number of possible endorsements ($n = 40$ for most of the analyses). This correction was applied 22 times across all participants, conditions, and tasks, most frequently on old endorsement rates that were at ceiling.

There were no interactions involving test list order or study list used so these variables will not be addressed further. An alpha level of .05 was used for all statistical tests.

--- Insert Table 1 about here ---

Structural and episodic discrimination. The endorsement rates in classification and recognition are shown in Table 1, whilst the d' measures are shown in Table 2. In addition to comparisons between the experimental conditions, the d' measures were compared against zero (chance performance) using the 95% confidence interval. These tests indicated that all d' s were significantly above zero (see Table 2).

--- Insert Table 2 about here ---

The d' measures were subjected to a 2 x 2 x 2 mixed ANOVA with within-subjects variables of task type (recognition/classification) and discrimination type (episodic/structural) and a between-subject variable of strength (weak/strong). Better overall discrimination was found in recognition ($M = 1.43$, $SE = .06$) than classification ($M = 1.17$, $SE = .06$), $F(1, 41) = 13.28$, $p = .001$, $\eta^2 = .25$ and in the strong condition ($M = 1.41$, $SE = .07$) compared to the weak condition ($M = 1.19$, $SE = .07$), $F(1, 41) = 4.99$, $p = .03$, $\eta^2 = .11$. In addition, d'_e was greater than d'_s , $F(1, 41) = 413.16$, $p < .001$, $\eta^2 = .91$.

More importantly, an interaction between strength and discrimination type, $F(1, 41) = 4.09$, $p = .05$, $\eta^2 = .09$, indicated that d'_e increased from weak to strong conditions, $F(1, 41) = 5.80$, $p = .02$, $\eta^2 = .13$, whereas d'_s did not, $F < 1$. There was also an interaction between task type and discrimination type, $F(1, 41) = 8.53$, $p = .006$, $\eta^2 = .18$ which indicated that d'_e decreased from recognition to classification $F(1, 41) = 13.67$, $p = .001$, $\eta^2 = .25$ whilst d'_s did not change by task, $F < 1$. See Table 2 for means and standard errors. There were no other significant effects, largest $F < 1$.

Awareness measures. The questionnaire indicated that no participant developed knowledge of the rule set. Even when directly told that the rule was a conjunctive rule, no participant accurately described the actual rule set. When forced to pick two items from a list of possible factors involved in the rule set, only one participant selected both frequency and concreteness, and the description he or she offered for how the two items were linked was incorrect. The most commonly selected individual items were *familiarity* and *frequency* (see Table 3). These results are analogous to those found by Neil and Higham (2012).

--- Insert Table 3 about here ---

Awareness was also tested with the zero-correlation criterion using Chan’s Difference Score (CDS; Dienes et al., 1995). CDS is calculated by subtracting mean confidence assigned to endorsed NI items from mean confidence assigned to endorsed NC items. If confidence for NC items is higher than that assigned to NI items, then CDS is larger than zero and explicit awareness is implicated. In particular, *judgment knowledge* (i.e., knowledge regarding whether or not an item has structure) is assumed to be conscious (Scott & Dienes, 2005). CDS scores were computed separately for classification and recognition and could range from 0 to 50 in either task.³ These CDS scores were compared against zero using 95% confidence intervals (see Table 4). The CDS scores were significantly above zero in all conditions except for the weak condition in the classification task. Variation in the CDS was investigated using a 2 x 2 mixed ANOVA with a within-subjects variable of task type (recognition/classification) and a between-subject variable of strength (weak/strong). There were no effects, highest $F(1,$

30) = 1.45, $p = .24$.⁴ However, it is worth noting that there was a numerical increase in the size of CDS as strength was increased, a point to which we return below.

--- Insert Table 4 about here ---

Finally, the distribution of participants' attributions to *random chance*, *intuition*, *memory* and *rules*, was examined to discover if the structural effect was limited to particular attributions or distributed more evenly. The endorsement rates were split by attribution type, and the difference between the endorsement rate for NC and NI items was calculated. This produced a difference score that indicated the amount of the structural effect that was attributed to the various attribution types (see Table 5). These difference scores were compared against zero (chance performance) using the 95% confidence interval. As shown in Table 5, the structural effect was attributed primarily to *intuition* and *memory* in both recognition and classification, with very little effect attributed to *random chance* or *rules*.

--- Insert Table 5 about here ---

Discussion

Consistent with previous results (Higham & Brooks, 1997; Neil & Higham, 2012), there was evidence in Experiment 1 of both a structural and an episodic effect in under classification instructions. The same was true under recognition instructions. There was, however, a dissociation between the episodic effect and the structural effect, as the increase in strength resulted in a greater increase in performance for the episodic effect than the structural effect. This finding is consistent with the extant theories and data. Were the task to be performed on the basis of rule-abstraction (e.g. Reber, 1967), then

responses to NC items would primarily be driven by the extent to which rule-abstraction was successful. Responding to old items would be supported by rule-abstraction plus the contribution of explicit memory for the old items, resulting in the old items having a greater sensitivity to strength than the NC items. If learning were based only on surface features, then the old stimuli are likely to contain more matching surface features than the NC items, again making the episodic effect increase more with strength than the structural effect. Instance-based models such as MINERVA and ATHENA (Curtis & Jamieson, 2019) can also explain the dissociation between the episodic and structural effects. As noted in the introduction, for old items responding is driven by the high similarity of the probe to a small number of individual instances in memory, whilst for NC items responding is driven by the weak similarity of the probe to a large number of instances in memory, and the former is more sensitive to strength manipulations than the latter. SRN (Kinder & Shanks, 2001) similarly suggests that the episodic and structural effects will be differentially sensitive to the strength manipulation.

More importantly, the structural effect did not increase with strength at all. This lack of increase was at odds with our predications derived from all of the theories mentioned above. Theories based on memory for surface features (e.g. Channon et al., 2002; Vokey & Higham, 2005) or rule abstraction (e.g. Reber, 1967), hold that the structural effect should have increased with exemplar repetition as long as test probes were, on average, more similar to NC than NI items.⁵ Whilst instance-based models suggest that the episodic and structural effects would be dissociated in their sensitivity to strength, repetition should still increase the structural effect, if only by a small amount. If this were happening, we should see at least a descriptive increase in the structural effect by strength, whereas descriptively the structural effect was actually higher in the weak condition in the strong condition, at least for in the recognition task. Of course, there may

1
2
3 be a small increase in the structural effect that we did not detect due to a lack of statistical
4
5 power. We will return to this point later, after Experiment 4. That said, in our view, our
6
7 finding of a static structural effect in the face of variable exemplar repetitions is
8
9 something that any theory of implicit learning needs to be able to explain.
10
11

12 In terms of awareness measures, Neil and Higham (2012) found that participants
13
14 were likely to experience the structural effect as familiarity, a finding that was replicated
15
16 in Experiment 1. Wan et al. (2008) suggested that participants can make strategic
17
18 decisions about how they use familiarity. However, the presence of a structural effect in
19
20 both classification and recognition suggests that no such strategic control was exercised
21
22 by our participants in Experiment 1. In the classification task, a structural effect is clearly
23
24 desirable (i.e., ideal performance would be $NC=1$; $NI=0$). Thus in the classification
25
26 condition, participants should strategically use familiarity to endorse NC items and reject
27
28 NI items, perhaps by endorsing all items with high familiarity. Conversely in the
29
30 recognition condition, no structural effect at all is desirable (i.e., ideal performance would
31
32 be $NC=NI=0$). To achieve this, participants would need to strategically use familiarity to
33
34 reject both NC and NI items, perhaps by requiring a higher level of familiarity to endorse
35
36 items compared to classification. Clearly participants have not successfully exercised
37
38 such strategic choices, instead endorsing items in such a way as to produce an equal
39
40 structural effect in both classification and recognition. Although Wan et al. have
41
42 demonstrated that participants can strategically use familiarity, our results suggest that
43
44 there is some element of familiarity that cannot be strategically controlled, or else that
45
46 some tasks result in participants making sub-optimal choices about their use of
47
48 familiarity.
49
50
51
52
53
54

55 The evidence from the post-experimental questionnaire suggested that no
56
57 participant gained verbalizable knowledge of the rule set. Even those participants who
58
59
60

selected the *rules* attribution were unable to accurately describe the rule set on the questionnaire. In contrast, the CDS analysis suggested that confidence in NC items was greater than NI items. Thus, on some level, participants’ subjective experience with NC items was different from that with NI items, which was reflected in their confidence judgments (i.e., the zero-correlation criterion was not met; Dienes et al., 1995). One possible reason for this difference between awareness measures is that participants, at least in part, included the higher familiarity of NC items compared to NI items as a basis for their confidence judgments. Nonetheless, while the post-experimental questionnaire results indicate no explicit awareness of the structural effect, the CDS analysis suggests that participants’ subjective confidence was sensitive to the structural manipulation, which is a form of awareness in its own right. Although not statistically significant, Table 4 suggests that strength may have descriptively increased CDS despite it having no effect on the item endorsement rates that underpinned the structural effect. We return to this point below.

The structural effect was distributed primarily across *intuition* and *memory* attribution types. If it is accepted that *intuition* and *memory* attributions represent unconscious and conscious structural knowledge, respectively (e.g., see Dienes & Scott, 2005), then the attribution data suggest both types of knowledge were present in Experiment 1. One potential reason that no structural effect was observed on *rules* attributions was that *rules* was the least popular attribution choice, particularly in recognition (see supplementary material). Nonetheless, the attribution results clearly indicate that the structural effect was not attributed to a single attribution type.

Experiment 2

Although Experiment 1 was successful in demonstrating that repetition has dissociative effects on the episodic versus structural effects in both recognition and classification, we conducted a second experiment with more challenging learning conditions. In particular, we shortened the display duration of the words during study. The reason for this change was that the endorsement rates of old items in both tasks in Experiment 1 were close to ceiling. The ceiling old endorsement rates necessitated applying a correction factor to multiple cells in the design to avoid undefined d' values, which could have potentially produced spurious results.

Second, participants in Experiment 1 were able to pause after reading each word during study whilst they made their understanding rating. This aspect of the procedure potentially allowed participants to rehearse each word such that the words were more or less perfectly encoded on first viewing. Under such circumstances, repeating the same exemplars may not have resulted in better learning. Therefore, in Experiment 2, we addressed this issue by removing the requirement for ratings during study.

One further change was implemented in Experiment 2 in an attempt to both remove potential confounding variables and to reduce the demands on participants. To avoid potential carry-over effects caused by requiring participants to engage in two test tasks, we manipulated task type between subjects in Experiment 2. This change was made because we reasoned that there might have been an influence of repetition on the structural effect in Experiment 1 that was masked by complexities introduced by requiring participants to perform both recognition and classification in different test blocks, and that a simpler design may allow such an influence to be revealed.

Method

Participants. Sixty-four undergraduates from the University of Southampton participated in the experiment for either course credits or £5 payment.

Materials and design. The materials and general design were the same as in Experiment 1.

Procedure. The procedure was similar to Experiment 1, except for the following changes. To address the ceiling effect for old-word endorsements, words in the study phase were displayed for 1 second each rather than 1.5 seconds (as in Experiment 1) and participants did not have to rate the words for understanding. Thus, whereas Experiment 1’s study phase allowed participants to rehearse each word before making their understanding rating, Experiment 2 allowed no such rehearsal time. Finally, only a single test block was administered to avoid any between-block carry-over effects. Thus, task type was a between- rather than a within-subjects variable in Experiment 2.

Results

Endorsement rates were calculated as in Experiment 1. As shown in Table 6, making the learning conditions more challenging in this experiment had the desired effect of reducing old-item endorsement rates that were near ceiling in Experiment 1. In fact, no corrections to the hit rate were required in this experiment (i.e., the old-item endorsement rate was never equal to one for any participant).

Eight participants were excluded from the analyses reported below – seven due to distractions from mobile phones occurring in the learning phase or guessing on every trial and one for very low recognition performance (more than three standard deviations below the mean).

--- Insert Table 6 about here ---

Structural and episodic discrimination. The d'_e and d'_s measures were again computed from the endorsement rates – see Table 6 for endorsement rates and Table 7 for d' measures. As in Experiment 1, the d' s were both compared against chance using confidence intervals (see Table 7) and the differences in the magnitude of the d' measures were investigated with a mixed 2 x 2 x 2 ANOVA with task type, strength (both between subjects) and effect type (within subjects).⁶

--- Insert Table 7 about here ---

As in Experiment 1, the ANOVA revealed that d'_e ($M = 1.14$, $SE = .06$) was higher than d'_s ($M = .24$, $SE = .04$), $F(1, 52) = 131.86$, $p < .001$, $\eta^2 = .72$. There was also better discrimination in the strong ($M = .81$, $SE = .05$) compared to the weak condition ($M = .57$, $SE = .05$), $F(1, 52) = 14.04$, $p < .001$, $\eta^2 = .90$, and in recognition ($M = .81$, $SE = .04$) compared to classification ($M = .57$, $SE = .05$), $F(1, 52) = 13.16$, $p < .001$, $\eta^2 = .20$.

Critically, there was an effect-type by study-strength interaction, $F(1, 52) = 8.51$, $p = .005$, $\eta^2 = .14$. Just as in Experiment 1, the interaction indicated an effect of strength on d'_e , $F(1, 52) = 17.07$, $p < .001$, $\eta^2 = .25$, but no change in d'_s , $F < 1$. There were no other significant effects from the ANOVA, largest $F < 1$.

Awareness measures. As in Experiment 1, the questionnaire data once again indicated that no participant developed explicit knowledge of the rule set. Only one participant selected the correct dimensions on the last question, but failed to describe the correct link between them. Contrary to Experiment 1, the most popular choices were *category (noun/verb)* and *association* (Table 3).

CDS was computed from the confidence data and compared against zero, as in Experiment 1 (Table 8). Only in the strong condition in the recognition task was the CDS different from zero, again indicating higher confidence for NC endorsements than NI endorsements. Variations in the CDS scores were analyzed with a 2 (task type: recognition/classification) x 2 (strength: strong, weak) mixed ANOVA. There was only a main effect of strength, $F(1, 52) = 4.78, p < .03, \eta^2 = .08$, indicating higher CDS scores in the strong condition compared to the weak condition. There were no other significant effects, highest $F(1, 52) = 2.09, p = 0.15, \eta^2 = 0.04$.

--- Insert Table 8 about here ---

As in Experiment 1, the difference scores between NC and NI endorsements were calculated for each attribution type (*random chance, intuition, memory* and *rules*), and compared against zero (see Table 9). Unlike Experiment 1, there was some evidence in classification that participants attributed the structural effect of *rules* instead of *intuition* and *memory*. However, *intuition* was the attribution source of the structural effect in recognition. Generally speaking, the attribution data indicate that there was both conscious and unconscious structural knowledge in Experiment 2, and that no one particular attribution was the source of the structural effect in either task.

--- Insert Table 9 about here ---

Discussion

As in Experiment 1, participants were sensitive to the episodic status of the items, and that sensitivity increased with strength. The strength effect was observed despite the poorer learning conditions in Experiment 2 compared to Experiment 1 (i.e., old endorsement rates such that they were no longer at ceiling). This effect suggests that the strength influence on the episodic effect is robust. More importantly, although a structural effect was observed in Experiment 2, it was not moderated by strength in either recognition or classification, replicating the results of Experiment 1.

As in Experiment 1, the questionnaire indicated that participants had no verbalizable knowledge of the rule set. However, unlike Experiment 1, the criteria that participants identified most commonly were *category (noun/verb)* and *association*. *Familiarity* was not a favored criterion, perhaps because the difficult learning conditions in Experiment 2 made any criterion to do specifically with memory influences less attractive.

Although the endorsement rates underpinning the structural effect were unaffected by strength, the strength manipulation reliably increased CDS. This result replicates a similar trend observed in Experiment 1. Thus, the results of Experiment 2 demonstrate that although strength does not moderate participants' tendency to *endorse* NC items more than NI items, it does moderate their *confidence* in those endorsements. Stated differently, exemplar repetition does not boost the structural effect in terms of endorsement rates, but it does alter the subjective confidence associated with those endorsements. As in Experiment 1, the fact that CDS was above zero (at least in the strong-recognition condition) and sensitive to strength suggests that the structural effect has an explicit component (Dienes & Scott, 2005), contrary to the post-experimental questionnaire results.

As in Experiment 1, the structural effect was spread across the different attribution sources (*random chance, intuition, memory, and rules*) rather than being associated with only one type. However, it is interesting to note that there was above-chance performance when participants said *rules* in classification, but there was no evidence that participants could actually verbalize the rule set used. Similar to Experiment 1, the fact that attributions were made to *intuition* on the one hand versus *rules* on the other, suggests that there was, respectively, both unconscious and conscious structural knowledge in Experiment 2.

Experiment 3

Both Experiments 1 and 2 demonstrated a dissociative influence of repetition on the episodic versus structural effects. Whereas the episodic effect increased dramatically with repetition of the rule-set exemplars during study, the structural effect remained invariant.

The general learning conditions were excellent in Experiment 1, which produced near ceiling endorsement rates for old items. This problem was alleviated in Experiment 2 by making the learning conditions more challenging. However, a separate issue pertains to the robustness of the strength manipulation. In both Experiments 1 and 2, strength was manipulated by presenting exemplars either once or thrice. Perhaps the structural effect was not sensitive to strength in those experiments because the strength manipulation was too weak. Conceivably, both the episodic and structural effects might be boosted with a more extreme manipulation of strength. To address this possibility, we used five presentations rather than three in the strong condition of Experiment 3.

A second possible issue with the strength manipulation is that the presence of strengthened items in the test list may have made the list as a whole seem more

memorable, resulting in some participants requiring a lot of evidence before responding “old” (recognition) or “consistent” (classification) during the test phase (i.e., a conservative responding strategy might have been adopted). However, if only some participants explicitly noticed the strength manipulation, then not all participants may have had the opportunity to adjust their decision criterion accordingly. To ensure that there were no such participant differences (which could have introduced noise and masked the moderating influence of strength on the structural effect), we made the strength manipulation in Experiment 3 more explicit by informing participants that they would see words during study either once or multiple times.

Method

Participants. Sixty-four undergraduates from the University of Southampton participated in the experiment either for course credits or £5 payment.

Materials and design. The materials and counterbalancing were the same as in Experiment 2, except for two stimuli that were replaced due to them sharing word fragments (e.g., *earth* and *earthworm*)⁷. The general design was also the same as Experiment 2 except for the following changes. To make the strength manipulation more noticeable, participants were warned, before the study phase began, about how many times each word would be presented. Additionally, words in the strong condition were displayed five times each (instead of three as in Experiments 1 and 2). A retention interval was also added between the study and test phases such that the time between the start of the study phase and the start of the test phase was ten minutes for every participant, regardless of the strength condition. During the retention interval, participants solved simple arithmetic problems in which a sum was presented with some digits missing (i.e. $45 + ?? = 59$). Participants were asked to write down the missing digits.

Note that this design feature meant that participants in the weak condition had a longer interval between seeing the last study word and the beginning of the test phase than did participants in the strong condition.

Procedure. The procedure was identical to Experiment 2, except for two aspects. The instructions were changed to inform participants how many times they would see each word during the study phase. Also, the retention interval described above was added.

Results

Mean endorsement rates in classification and recognition were calculated and corrected as previously (six corrections were applied) and are shown in Table 10. One participant was excluded from the analysis for low recognition performance (more than three standard deviations below the mean).

--- Insert Table 10 about here ---

Structural and episodic discrimination. The mean d'_e and d'_s measures in classification and recognition were again computed from the endorsement rates for each participant, see Table 11. Mean d 's were then compared to chance using confidence intervals and differences in magnitude of the d' measures were investigated with a 2 x 2 x 2 ANOVA with between subject variables of task type and strength and a within-subject variable of effect type.

--- Insert Table 11 about here ---

The ANOVA again revealed that there was a higher d'_e ($M = 1.29$, $SE = 0.11$) than d'_s ($M = 0.32$, $SE = 0.04$), $F(1, 59) = 73.66$, $p = .001$, $\eta^2 = .55$ and there was overall better

discrimination in the strong ($M = 1.08$, $SE = 0.07$) than the weak ($M = 0.53$, $SE = 0.07$) condition, $F(1, 59) = 38.54$, $p < .001$, $\eta^2 = .85$. Critically, as in Experiments 1 and 2, there was an interaction between effect type and strength, $F(1, 59) = 41.11$, $p < .001$, $\eta^2 = .29$. As previously, the interaction reflected a greater d'_e in the strong than in the weak condition, $F(1, 59) = 38.45$, $p < .001$, $\eta^2 = .39$, whilst the d'_s did not change, $F < 1$ (Table 11). There were no other significant effects, highest $F(1, 59) = 1.07$, $p = .30$.

Awareness questionnaire. Once again the questionnaire indicated that no participants developed explicit knowledge of the rule set. No participant guessed the two correct criteria, with the most popular selections being *association*, *frequency*, and *meaning* (Table 3). Thus, unlike Experiment 1, but similar to Experiment 2, *familiarity* was not the favorite criterion amongst participants.

CDSs were analyzed as in the previous experiments (see Table 12). For both tasks, the CDS scores in the strong condition were significantly above zero whilst the CDS scores in the weak condition were not. As in Experiment 2, a 2 (task type: recognition/classification) \times 2 (strength: strong, weak) mixed ANOVA indicated only a main effect of strength, $F(1, 59) = 5.85$, $p = .02$, $\eta^2 = .09$, reflecting higher CDS scores in the strong condition than in the weak condition.

--- Insert Table 12 about here ---

The difference scores between NC and NI endorsements for each attribution type (*random chance*, *rules*, *memory*, and *intuition*) were also analyzed (see Table 13). As in previous experiments, there was no evidence that one particular attribution was the source of the structural effect. Instead, *memory*, *intuition* and *rules* all played a part. Thus, as in

the previous experiments, there was evidence of both conscious (*memory, rules*) and unconscious (*intuition*) structural knowledge in this experiment.

--- Insert Table 13 about here ---

Discussion

As in the previous two experiments, repetition increased the episodic effect, but had no effect on the structural effect. This occurred despite the fact that the repetition manipulation was boosted from 1 vs. 3 repetitions in Experiments 1 and 2 to 1 vs. 5 in Experiment 3. Consequently, there is no evidence that the dissociation observed in Experiments 1 and 2 was due to a weak strength manipulation. Instead, it appears that the structural effect, unlike the episodic effect, is insensitive to strength under a variety of experimental conditions.

As in the previous experiments, the questionnaire data indicated that participants did not acquire verbalizable knowledge of the rules. This was true despite presenting participants, toward the end of the questionnaire, with the exact categories that made up the conjunctive rule set. Unlike Experiment 1, *familiarity* was not participants’ favorite criterion, just as it was not in Experiment 2. The *d’* data suggest that the learning conditions in Experiment 3 were similar to Experiment 2 and not as favorable as they were in Experiment 1 (cf. Tables 2, 7 and 11). These poorer conditions in the latter two experiments may have dissuaded participants from selecting *familiarity* as their judgment criterion, as familiarity is a memory-based attribution and memory for the items would play a less significant role with poorer learning conditions.

In contrast to the endorsement data, the confidence data indicated that the CDS increased with exemplar repetition in both tasks, which meant that they were above chance in the strong condition, but not the weak condition. In the strong condition (but less so in the weak condition) participants endorsed NC words with higher confidence than NI words, regardless of whether it was appropriate (classification) or inappropriate (recognition) to do so. Thus, whereas the endorsement rates underpinning the structural effect in both classification and recognition were unaffected by strength, there was evidence that subjective confidence responded to the strength manipulation. However, our post-experimental questionnaire results suggest that any changes to subjective confidence arising from strength did not translate into verbalizable knowledge.

As in Experiments 1 and 2, the structural effect was distributed across the different attribution types rather than being limited to one attribution in particular. Even though there was again some evidence of a structural effect when participants selected the *rules* attribution, participants were not able to verbalize what this rule set might have been.

Experiment 4

Across Experiments 1-3, strength did not increase the structural effect. One possible explanation for this result is that, as additional exemplars leads to better learning (Reber, 1989), 80 exemplars is enough for participants to extract all possible information about the ruleset. Consequently, additional presentations of the same exemplars provided no learning benefit. Thus, to address this issue, Experiment 4 introduced a condition with 5 exemplars per ruleset, for a total of 10 exemplars at study. If in the previous experiment participants had simply reached the limit of their learning, then at such a low number of exemplars, strength may begin to induce an increase in the structural effect

Method

Participants. One hundred and sixty-two undergraduates from the University of Southampton participated in the experiment either for course credits or £5 payment.

Materials and design. The materials and counterbalancing were the same as in Experiment 3. The general design was also the same as Experiment 3 except for an additional variable. As well as study strength, variety at study was also manipulated, between participants. The study phase contained either 80 unique words (as in Experiment 3) or 10 unique words. In addition, in order to focus on the structural effect, only a classification task was used, and old words were not displayed at test. Thus, the test words were all NC or NI, and all participants took part in the classification task.

Procedure. The procedure was identical to Experiment 3, other than the additional condition noted above, and the lack of a recognition task.

Results

Mean endorsement rates are shown in Table 14. It was not necessary to correct any rates of 0 or 1.

--- Insert Table 14 about here ---

Structural and episodic discrimination. The mean d' s were again computed from the endorsement rates for each participant and are shown in Table 15. Mean d' s were then compared to chance using confidence intervals and differences in magnitude of the d' measures were investigated with a 2 x 2 ANOVA with between-subject variables of variety and strength.

--- Insert Table 15 about here ---

For both the weak and strong conditions, all $d's$ were above chance - see Table 15. The ANOVA revealed only an effect of variety - $d's$ was higher in the 80 word condition than the 10 word condition, $F(1, 158) = 4.00, p = .047, \eta^2 = .025$. There were no other effects, highest $F(1, 158) = 1.44, p = 0.23$.

Awareness questionnaire. Once again the questionnaire indicated that no participants developed explicit knowledge of the rule set. No participant guessed the two correct criteria, with the most popular selections being association, category, and meaning (Table 3). Thus, as in Experiments 2 and 3, familiarity was not the favorite criterion amongst participants.

CDS measures were analyzed as in the previous experiments (see Table 16). For both tasks, the CDS scores in the strong condition were significantly above zero whilst the CDS scores in the weak condition were not. Changes in the CDS measures were analyzed using a 2 (variety: 10 words/80 words) x 2 (strength: strong, weak) ANOVA. There was only a main effect of strength, $F(1, 158) = 10.98, p = .001, \eta^2 = .06$, reflecting higher CDS scores in the strong condition than in the weak condition. There were no other effects, highest $F < 1$.

--- Insert Table 16 about here ---

As in the previous experiments, the difference scores between NC and NI endorsements were calculated for each attribution type, and compared against zero (see Table 17). Across the various conditions, random chance, memory and rules were all implicated. Once again, there seemed to be both conscious and unconscious structural knowledge, with no one attribution being dominant.

--- Insert Table 17 about here ---

Discussion

With both 10 and 80 exemplars at study, the results of Experiment 4 were similar to Experiment 3. The structural effect did not increase with strength, but CDS ratings did. However, the structural effect did increase with the number of exemplars used, with the structural effect being greater in the 80 words condition than in the 10 words condition. Thus, it would seem that the lack of increase in the structural effect with strength was not due to the structural effect being at ceiling, otherwise it should have increased with strength in the 10 word condition. Whilst this result supports the finding of Reber (1989), it is also similar to work in the category learning literature, in which increased exemplar variety also leads to increased performance (Dukes & Bevan, 1967; Homa, Sterling, & Trepel, 1981). Regardless, it seems that even at low numbers of exemplars, repetition did not help increase the magnitude of the structural effect.

Pooled Analysis

Across four experiments, the structural effect was invariant to exemplar repetition whilst the episodic effect greatly increased with repetition. However, it may be that the previous experiments simply did not have sufficient power to detect a change in the structural effect. In this section, we conduct additional analyses to discount this explanation.

Although the structural effect was statistically reliable in the majority of conditions across all reported experiments, the associated effect size was much smaller than that associated with the episodic effect. This result suggests that the structural effect was small and noisy, which may have masked the moderating influence of repetition.

Furthermore, the structural effect may have been unaffected by repetition because some participants in the full sample failed to make a structural discrimination at all (i.e., NC endorsements = NI endorsements). These participants may have served to mask any effects of repetition in the rest of the sample.

Two additional analyses were conducted to investigate these possibilities. The first analysis combined the data from all four experiments. This ensured that the analysis had more power than the individual experiment analyses. The second analysis selectively analyzed participants from all four experiments who demonstrated a structural effect that was in the 67th percentile or higher (i.e., structural effects that were in the top third). This analysis would thus exclude any participants who failed to make a structural discrimination.

Neither analysis changed the general conclusion. For each analysis a 4 (Experiment: 1, 2, 3, 4) X 2 (strength: weak, strong) X 2 (task: recognition, classification) mixed ANOVA was conducted on the d'_s measures. There were no effects of strength in either the ANOVA including all participants, weak d'_s $M = 0.26$ ($SE = 0.04$), strong $d'_s = 0.25$ ($SE = 0.04$), highest $F(1, 209) = 0.70, p = .55$. Nor was there an effect of strength for the 67th percentile analysis, weak d'_s $M = 0.63$ ($SE = 0.05$), strong $d'_s = 0.60$ ($SE = 0.05$), highest $F(1, 65) = 1.50, p = .23$. Thus, even if the analysis included all participants across all four experiments or was limited to only those participants who were particularly successful at making the structural discrimination, repetition did not increase the structural effect. The results of this analysis strongly support the original conclusions of the above experiments that the structural effect does not increase from the weak to the strong conditions. The lack of sensitivity of the structural effect to strength was thus not due to a lack of power in the experiments.

To further investigate whether the data represented an undetected change in the structural effect, a Bayesian analysis of the data was conducted. Bayesian analyses have the advantage that they can be used to draw inferences about whether there is evidence for the null hypothesis or not (see Jarosz & Wiley, 2014 for more information on Bayesian statistics). Thus, for both all participants and the 67th and above percentile participants, the data were collapsed across all variables other than strength, and entered into a Bayesian T-test using the JASP v0.9 (REF – JASP Team 2018), with the alternate hypothesis set to be that the weak condition was not the same as the strong condition. The measure of BF_{01} was used, which indicates how much more likely the data was if the null were true than if the alternate were true. For all participants, the $BF_{01} = 6.47$, and for the 67th percentile participants the $BF_{01} = 3.23$. According to Jeffreys (1961) this constitutes “substantial” evidence that the data we observed reflected no actual change in the structural effect between weak and strong conditions.

General Discussion

Across four experiments, we investigated the effect that exemplar repetition had on both discrimination between old versus NC items (the episodic effect) and between NC versus NI items (the structural effect). Regardless of whether exemplars were repeated three times (Experiments 1 and 2) or five times (Experiments 3 and 4), repetition had dissociative effects on the two forms of discrimination. In particular, repeating the exemplars greatly boosted the episodic effect, but left the structural effect invariant. This dissociation occurred regardless of whether the rule set was represented by 80 or by 10 different exemplars (Experiment 4). Hence, the dissociation cannot be explained by the rule set being already fully learned, as the structural effect still did not increase by

1
2
3 strength even when it was represented by a smaller number of exemplars. Explanations
4
5 based on a lack of power or a noisy structural effect were investigated and discounted.
6

7
8 Dissociations analogous to these have been reported previously in the literature,
9
10 but usually as task dissociations. For instance, SRN (Kinder & Shanks, 2001) and
11
12 MINERVA variants such as ATHENA (Curtis & Jamieson, 2019) have been used to
13
14 demonstrate dissociations in study-item strength effects for old/new discrimination in
15
16 recognition compared to NC/NI classification (see also Knowlton et al., 1992). Our
17
18 research differs from that previous research in that we demonstrated the dissociation
19
20 *within* both the recognition and classification tasks by comparing different items types.
21
22 Most models of implicit learning can account for this dissociation. For example, the
23
24 dissociation is predicted by MINERVA because the extent to which memory traces are
25
26 activated by a probe is a cubic function of the similarity between the traces and the probe.
27
28 This function produces a non-linear generalization gradient such that highly similar traces
29
30 (e.g., old items) have much greater influence than weakly or moderately similar traces
31
32 (e.g., NC items). Hence, strengthening old items (e.g., through repetition or greater
33
34 presentation time) will have a much greater effect on old-item endorsements than NC-
35
36 item endorsements, whereas any influence on NI-item endorsements will be negligible.
37
38
39
40
41

42 This discussion makes clear that probe/trace similarity is critical to understanding
43
44 the effect that study strength manipulations will have on the episodic and structural effects
45
46 within MINERVA. If the NC items were highly similar to memory traces, then the effect
47
48 of strength on the structural effect would approach that of the episodic effect. However, it
49
50 is important to point out that even weak NC/trace similarity should be responsive to
51
52 strength under most circumstances. Indeed, current simulations using instance-based
53
54 models suggest that whereas dissociations such as those found in our experiments do
55
56 occur, classification performance still improves to some degree with learning strength
57
58
59
60

(e.g., Curtis & Jamieson, 2019). This is true not just of MINERVA, but of other theoretical accounts of implicit learning as well. For example, if learning is based on unconscious rule abstraction (e.g. Reber, 1989), then repeated exposure to exemplars provides more opportunities for that abstraction to occur. Thus, a central question posed by our results is why the structural effect was not sensitive to study strength whilst the episodic effect was.

This question is highly relevant given that strength-based enhancement of the structural effect should have been particularly likely in our experiments given that the rule set was not instantiated with just a few exemplars, but with many different exemplars. In what follows, we will consider some potential theoretical accounts as to why the structural effect was completely insensitive to strength.

Accounting for the Data

Differentiation. A possible route to explaining why the structural effect was insensitive to strength is through the concept of *differentiation*. Differentiation has been discussed in both the implicit learning and recognition memory literatures. For example, Vokey and Brooks (1992) gave participants unique mnemonics to assist in encoding AG strings (e.g., *MVX* was associated with *Mandy Viewed X-rays*). They found if the items were individuated in this way, transfer of learning to new items was reduced. This led Vokey and Brooks to conclude:

“...learning conditions that result in enhanced retrieval of specific item information can decrease the effectiveness of such information to support generalization. Apparently, items can be too well differentiated to support effectively transfer performance on new items (p. 338).”

Similarly, the concept of differentiation has been formalized into two models of memory: Shiffrin and Steyvers' (1997) REM model, and McClelland and Chappell's (1998) Subjective Likelihood Model (SLiM). In both cases, good learning of individual exemplars in the study phase of either an implicit learning or recognition memory experiment causes those exemplars to become individuated and differentiated from other exemplars. In other words, well-learned exemplars become unique and dissimilar to other exemplars because the differences that exist between the well-learned exemplars and the other exemplars are also well learned.

Another example of a model that uses differentiation is the network-based SRN model. One important feature of this model is that it can produce data consistent with our findings. For example, SRN simulations produce either no increase or even a decrease in NC/NI discrimination with increases in memory strength (Kinder & Shanks, 2001). Critically, however, SRNs specifically model *sequential* relationships such as those inherent in AG stimuli. The stimuli used in our experiment used more naturalistic, semantic categories of words to define the rule-set, and such stimuli exhibit different properties to those used in an AG task (Neil & Higham, 2012). Consequently, it is not clear how SRN and similar sequential models would represent the underlying semantic information needed to simulate the structural effect in the first place.

Nevertheless, in our own experiments, it is possible differentiation may have occurred, resulting in opposing influences acting on repeated NC items in the test phase. On the one hand, the repetitions allowed for multiple opportunities to learn the *commonalties* amongst the studied exemplars, which should support good transfer. On the other hand, repetition allowed for more *differentiation*, which would limit transfer. Assuming these opposing influences were approximately equal, the net result would be a

structural effect that was invariant as the number of repetitions increased, exactly the result that we observed.

Although enticing, this explanation may be a little too convenient. It entails that these two opposing influences would have to perfectly balance each other across the varying conditions of Experiments 1 to 4, including different numbers of repetitions, different study and test conditions, and different task types. Consequently, a sufficient explanation for our results would likely need to involve more than just the learning of differences and commonality cancelling each other out.

Processes and representations. In assessing the reason that the structural effect was invariant to repetition, some consideration of what participants might actually be learning is warranted. However, as with most implicit learning experiments, identifying the exact source of learning is not necessarily straightforward. For example, in AGL research, several researchers have argued that participants do not learn the underlying finite state grammar, but rather respond to the frequency of bigram and trigram chunks in the studied items (e.g. Dulany et al., 1984; Perruchet & Pacton, 2006; although see Higham, 1997a;). However, Neil & Higham (2012), using materials similar to those used here, demonstrated that something other than chunk learning was the basis of the structural effect in their research. Nonetheless, the possibility exists that, just as with AGL research, conjunctive rule-set learning involves sensitivity to something *correlated* with the conjunctive rule set rather than the rule set itself.

For example, instead of learning that the words were either concrete or abstract, participants might learn that the words are either detectable by the one of the five senses or not (e.g., box vs love). Such learning is likely correlated with abstractness and concreteness, and thus would still produce a structural effect (as long as this dimension was learned in conjunction with lexical frequency or something correlated with lexical

frequency). Another possibility is that the specific *meaning* of the individual words (rather than their concreteness/abstractness or lexical frequency) is somehow correlated with the conjunctive rule set. Indeed, Chubala (2016) showed that the structural effect observed in Neil and Higham (2012) could be simulated using MINERVA 2 if the vectors were derived from semantic space models such as BEAGLE or LSA.

Critically, however, the importance of our current findings is not dependent on precisely ascertaining what was learned. Regardless of whether participants are learning the actual rule set or something correlated with it, repetition should have increased the structural effect simply because repetition affords more opportunities for learning. However, this increase was not observed in any of our experiments.

An added assumption that could potentially explain the strength-based invariance of the structural effect would be that repeated words were processed poorly, adding very little to what was learned on the first presentation. There is some evidence from various literatures that such impaired processing of repetitions may occur. Runger and Frensch (2008) demonstrated that expected events are processed less than unexpected events, whilst Hintzman and Curran (1995) demonstrated similar poor processing of repeated items with their *registration-without-learning* phenomenon. With this phenomenon, wastage of processing resources was prevented by only fully analyzing novel stimuli repeated stimuli were processed only for frequency of occurrence. Similarly, in the context of the spacing effect (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), *deficient processing theory* (e.g., Hintzman, 1974) has been offered as an explanation for why massed repetitions are not learned very well. Finally, a similar idea exists in associative learning in the form of latent inhibition, again where unsurprising stimuli no longer result in associations being formed (Pearce & Hall, 1980). If the processing of repeated words in our experiments was deficient for some reason, then learning may have

1
2
3 been impaired, and the structural effect would not increase with strength. Strength-based
4
5 invariance of the structural effect would occur under these circumstances regardless of
6
7 whether the actual rule set or something correlated with it was being. Models of learning
8
9 and memory could be adapted to simulate such degraded processing of repeated items.
10
11 For instance, repeated stimuli could be assigned a lower L than non-repeated stimuli in
12
13 MINERVA, although it would need to be done in such a way that the episodic effect still
14
15 increased with strength. Whether such degraded processing of repeated stimuli occurred
16
17 here or not is an empirical question which could be investigated in future research.
18
19
20

21 **Awareness.** As we have noted, none of our participants were able to accurately
22
23 describe the rules on a post-experimental questionnaire, even after being presented with
24
25 the specific criteria that made up the rule set. However, to further understand our results
26
27 in terms of awareness, we believe it may be fruitful to draw on Dienes and Scott’s (2005)
28
29 distinction between *structural* knowledge and *judgment* knowledge. According to their
30
31 analysis, structural knowledge can be either conscious or unconscious, and the status of
32
33 structural knowledge can be ascertained by analyzing the *memory*, *rules*, *intuition*, and
34
35 *guess/random chance* attributions, the same attributions we used in our research.
36
37 Specifically, if attributions to *memory* or *rules* show a structural effect, then structural
38
39 knowledge is said to be conscious. Conversely, if attributions to *guess/random chance* or
40
41 *intuition* show a structural effect, then structural knowledge is said to be unconscious.
42
43 The fact that we obtained significant structural effects at different points across our tasks
44
45 and experiments with *intuition*, *rules* and *memory* attributions suggests that some portion
46
47 of our participants’ responses constituted conscious structural knowledge (i.e., structural
48
49 effect for either *memory* or *rules*) and some portion constituted unconscious structural
50
51 knowledge (structural effect for *intuition*). However, it is interesting to note that even for
52
53
54
55
56
57
58
59
60

items for which there was conscious structural knowledge by Dienes and Scott's (2005) analysis, participants were still unable to verbalize the structure.

In the experiments above, we measured judgement knowledge using the CDS measure. According to Dienes and Scott (2005), unconscious structural knowledge can be associated with either conscious or unconscious judgment knowledge. For example, *intuition* attributions yielding a structural effect (unconscious structural knowledge) may have a CDS above chance (conscious judgment knowledge) or a CDS at chance (unconscious judgment knowledge), and the same is true of *guess/random chance* attributions. Although we did not examine specifically CDS as a function of attribution type, it is worth noting that CDS scores were above chance in several conditions across the experiments, which suggests that judgment knowledge was conscious. Furthermore, the CDS scores increased with strength whereas the structural effect (i.e., structural knowledge) did not. Hence, if one uses CDS to measure judgment knowledge as distinct from structural knowledge, then the dissociative pattern of endorsements versus confidence ratings that we observed in our experiments has a theoretical foundation. Specifically, our results suggest that exemplar repetition does not affect structural knowledge (conscious or unconscious) but it does affect the acquisition of conscious judgment knowledge. In our view, it is important to investigate this possibility in future research.

Final Thoughts

The fact that the recognition and classification tasks produced similar data demonstrates that there should be closer links between recognition and implicit-learning literatures. A link between classification and recognition is not a new concept – for instance, Higham (1997b; Higham & Brooks, 1997) as well as others (e.g., Vokey & Brooks, 1992; Whittlesea & Dorken, 1993) have pointed out the fact that both tasks seem

to be performed using the same processes. In fact, in the context of implicit learning theories, it would seem to be difficult not to talk about recognition memory at the same time as classification. Study of the differences between the two task types could also be revealing, and computational models will need to account for how performance may differ in response to task demands.

That the structural effect did not increase with strength over four experiments is a surprising result. Although we have speculated as to why this may be the case, future research needs to put these speculations to the test. For now, the unresponsiveness of the structural effect to repetition remains a puzzle that models of memory and learning will need to solve.

Declaration of Conflicting Interests

The authors declare that there are no conflicts of interest.

Funding Acknowledgements

This research was supported in part by an ESRC studentship ES/F02200x/1 to the first author.

Supplementary Material

The Supplementary Material is available at: qjep.sagepub.com

References

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization* (pp. 3-170): Lawrence Elbaum Associates.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354-380. doi:10.1037/0033-2909.132.3.354

Channon, S., Shanks, D., Johnstone, T., Vakili, K., Chin, J., & Sinclair, E. (2002). Is implicit learning spared in amnesia? Rule abstraction and item familiarity in artificial grammar learning. *Neuropsychologia*, 40(12), 2185-2197. doi:10.1016/S0028-3932(02)00037-4

Chubala, C. M., Johns, B. T., Jamieson, R. K., & Mewhort, D. J. K. (2016). Applying an exemplar model to an implicit rule-learning task: Implicit learning of semantic structure. *The Quarterly Journal of Experimental Psychology*, 69(6), 1049-1055. doi:10.1080/17470218.2015.1130068

Coltheart, M. (1981). MRC Psycholinguistic database user manual: version 1.

Curtis, E. T., & Jamieson, R. K. (2019). Computational and empirical simulations of selective memory impairments: Converging evidence for a single-system account of memory dissociations. *Quarterly Journal of Experimental Psychology*, 72(4), 798-817. doi:10.1177/1747021818768502

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5), 1322-1338. doi:10.1037/0278-7393.21.5.1322

- Dienes, Z., Broadbent, D., & Berry, D. C. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(5), 875-887. doi:10.1037/0278-7393.17.5.875
- Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research/Psychologische Forschung*, 69(5-6), 338-351. doi:10.1007/s00426-004-0208-3
- Dukes, W. F., & Bevan, W. (1967). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology*, 74(2p1), 178.
- Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General*, 113(4), 541-555. doi:10.1037/0096-3445.113.4.541
- Frensch, P. A. (1998). One concept, multiple meanings: On how to define the concept of implicit learning. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning*. (pp. 47-104). Thousand Oaks, CA US: Sage Publications, Inc.
- Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91(1), 1-67.
- Higham, P. A. (1997a). Chunks are not enough: The insufficiency of feature frequency-based explanations of artificial grammar learning. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Experimentale*, 51(2), 126-138. doi:10.1037/1196-1961.51.2.126
- Higham, P. A. (1997b). Dissociations of grammaticality and specific similarity effects in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(4), 1029-1045. doi:10.1037/0278-7393.23.4.1029

- Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. *The Quarterly of Experimental Psychology*, 50A(1). doi:10.1080/027249897392297
- Higham, P. A., Bruno, D., & Perfect, T. J. (2010). Effects of study list composition on the word frequency effect and metacognitive attributions in recognition memory. *Memory*, 18(8), 883-899. doi:10.1080/09658211.2010.517757
- Higham, P. A., & Vokey, J. R. (1994). Recourse to stored exemplars is not necessarily explicit: A comment on Knowlton, Ramus, and Squire (1992). *Psychological Science*, 5(1), 59-60. doi:10.1111/j.1467-9280.1994.tb00617.x
- Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In *Theories in cognitive psychology: The Loyola Symposium*. (pp. xi, 386-xi, 386). Oxford, England: Lawrence Erlbaum.
- Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers*, 16(2), 96-101. doi:10.3758/BF03202365
- Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, 93(4), 411-428. doi:10.1037/0033-295X.93.4.411
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95(4), 528-551. doi:10.1037/0033-295x.95.4.528
- Hintzman, D. L., & Curran, T. (1995). When encoding fails: Instructions, feedback, and registration without learning. *Memory & Cognition*, 23(2), 213-226.
- Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, 7(6), 418.

- Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *The Quarterly Journal of Experimental Psychology*, 62(3), 550-575.
doi:10.1080/17470210802055749
- Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction-time task: Anticipating from experience. *The Quarterly Journal of Experimental Psychology*, 62(9), 1757-1783. doi:10.1080/17470210802557637
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting bayes factors. *Journal of Problem Solving*, 7(1), 2-9.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.
- Jiang, S., Zhu, L., Guo, X., Ma, W., Yang, Z., & Dienes, Z. (2012). Unconscious structural knowledge of tonal symmetry: Tang poetry redefines limits of implicit learning. *Consciousness and Cognition: An International Journal*, 21(1), 476-486.
doi:10.1016/j.concog.2011.12.009
- Kinder, A., & Shanks, D. R. (2001). Amnesia and the declarative/nondeclarative distinction: A recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience*, 13(5), 648-669.
doi:10.1162/089892901750363217
- Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for specific instances. *Psychological Science*, 3(3), 172-179.
- Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1), 79-91. doi:10.1037/0278-7393.20.1.79

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present day american english*. Providence: Brown university press.

Kuhn, G., & Dienes, Z. (2005). Implicit learning of nonlocal musical rules: Implicitly learning more than chunks. *Journal of Experimental Psychology-Learning Memory and Cognition*, 31(6), 1417-1432. doi:10.1037/0278-7393.31.6.1417

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, 52(3), 436-459.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105(4), 724-760. doi:10.1037/0033-295X.105.4.734-760

Meulemans, T., & Van der Linden, M. (1997). Associative chunk strength in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(4), 1007-1028. doi:10.1037/0278-7393.23.4.1007

Neil, G. J., & Higham, P. A. (2012). Implicit learning of conjunctive rule sets: An alternative to artificial grammars. *Consciousness and Cognition: An International Journal*. doi:10.1016/j.concog.2012.07.005

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19(1), 1-32. doi:10.1016/0010-0285(87)90002-8

Norman, E., Price, M. C., Duff, S. C., & Mentzoni, R. A. (2007). Gradations of awareness in a modified sequence learning task. *Consciousness and Cognition: An International Journal*, 16(4), 809-837. doi:10.1016/j.concog.2007.02.004

- Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87(6), 532-552. doi:10.1037/0033-295x.87.6.532
- Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, 119(3), 264-275. doi:10.1037/0096-3445.119.3.264
- Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, 10(5), 233-238.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, 6(6), 855-863. doi:10.1016/S0022-5371(67)80149-X
- Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118(3), 219-235. doi:10.1037/0096-3445.118.3.219
- Rowland, L. A., & Shanks, D. R. (2006). Sequence learning and selection difficulty. *Journal of Experimental Psychology-Human Perception and Performance*, 32(2), 287-299.
- Runger, D., & Frensch, P. A. (2008). How incidental sequence learning creates reportable knowledge: The role of unexpected events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1011-1026. doi:10.1037/a0012942
- Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1264-1288. doi:10.1037/a0012943
- Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4), 592-608. doi:10.1037/0278-7393.16.4.592

- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2), 145-166.
- Song, S., Howard, J. H., Jr., & Howard, D. V. (2007). Implicit probabilistic sequence learning is independent of explicit awareness. *Learning & Memory*, 14(3), 167-176. doi:10.1101/lm.437407
- Song, S., Howard, J. H., Jr., & Howard, D. V. (2008). Perceptual sequence learning in a serial reaction time task. *Experimental Brain Research*, 189(2), 145-158.
- Tunney, R. J. (2007). The subjective experience of remembering in artificial grammar learning. *European Journal of Cognitive Psychology*, 19(6), 934-952. doi:10.1080/09541440600778083
- Tunney, R. J. (2010). Similarity and confidence in artificial grammar learning. *Experimental Psychology*, 57(2), 160-168. doi:10.1027/1618-3169/a000020
- Van den Bos, E., & Poletiek, F. H. (2008). Effects of grammar complexity on artificial grammar learning. *Memory & Cognition*, 36(6), 1122-1131. doi:10.3758/MC.36.6.1122
- Vokey, J. R., & Brooks, L. R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2), 328-344. doi:10.1037/0278-7393.18.2.328
- Vokey, J. R., & Higham, P. A. (2005). Abstract analogies and positive transfer in artificial grammar learning. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie experimentale*, 59(1), 54-61. doi:http://dx.doi.org/10.1016/S1196-1961(05)60239-5
- Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition*, 40(5), 703-716. doi:10.3758/s13421-011-0180-2

Wan, L., Dienes, Z., & Fu, X. (2008). Intentional control based on familiarity in artificial grammar learning. *Consciousness and Cognition: An International Journal*, 17(4), 1209-1218. doi:10.1016/j.concog.2008.06.007

Whittlesea, B. W. A., & Dorken, M. D. (1993). Incidentally, things in general are particularly determined - an episodic-processing account of implicit learning. *Journal of Experimental Psychology-General*, 122(2), 227-248. doi:10.1037/0096-3445.122.2.227

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, Version 2.00. *Behavior Research Methods, Instruments & Computers*, 20(1), 6-10.

Footnotes

¹ The term *guess* has several meanings and is therefore prone to different interpretations. To avoid this confusion in our participants, the term *guess* was replaced with *random chance* – see Scott and Dienes (2008) Experiment 3.

² d'_e was calculated by subtracting the z transformed NC endorsement rate from the z old endorsement rate. d'_s was calculated by subtracting the z NI rate from the z NC rate. Traditional signal-detection theory calculates d' using hit and false-alarm rates. However, since the NC rate is a false-alarm rate in recognition and a hit rate in classification this terminology has been avoided.

³ In previous research (e.g. Dienes et al., 1995) CDS was computed for NC items versus NI items in classification rather than recognition. In classification, endorsing the former item type constitutes a correct response whereas endorsing the latter constitutes an error. As such, CDS becomes a measure of *metacognitive resolution*, the degree to which people can monitor the correctness of their responding. In recognition, however, endorsement of any new item constitutes an error, regardless of whether the item is consistent or inconsistent with the rule structure. Therefore, in recognition, any difference in confidence is not a measure of metacognitive resolution per se. However, although the meaning of CDS is somewhat different in classification versus recognition, it is still useful for assessing the subjective experience associated with endorsing the different stimulus types and can still be used to index awareness: if confidence is higher for endorsements of NC versus NI items, then the rule structure is affecting participants' subjective experience, which is a form of awareness. This conclusion is true regardless of whether the task is classification or recognition.

⁴ Reduced degrees of freedom in the analysis were due to cases dropped with no endorsements for either NC or NI stimuli

⁵ As noted above, this condition may not hold for theories of implicit learning based on memory for surface features with our materials because chunk frequency is constant between NC and NI items (Neil & Higham, 2012).

⁶ Careful readers may note that the structural effect for classification was smaller than for recognition and was not significantly above zero in either the weak or strong conditions. However, it is worth making two points in this regard. First, the structural effect in classification was significantly above

zero if the weak and strong conditions were pooled (see Table 7), suggesting the failure to find a significant effect in either the weak or strong condition alone was a power issue. Second, the structural effects in both Experiments 1 and 3 were comparable between classification and recognition, suggesting that the larger effect observed in Experiment 2 for recognition is not reliable.

⁷ The materials used in this experiment can be found in the supplementary materials.

Figure Captions

Figure 1 – Signal detection model for Experiments 1 to 3. Participants are assumed to rely on a yes/no criterion (C) applied to the strength-of-evidence dimension to make both recognition and classification judgments. Old = words seen at study, new = words not seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words, d'_s = Structural signal detection measure d' , d'_e = episodic signal detection measure d' .

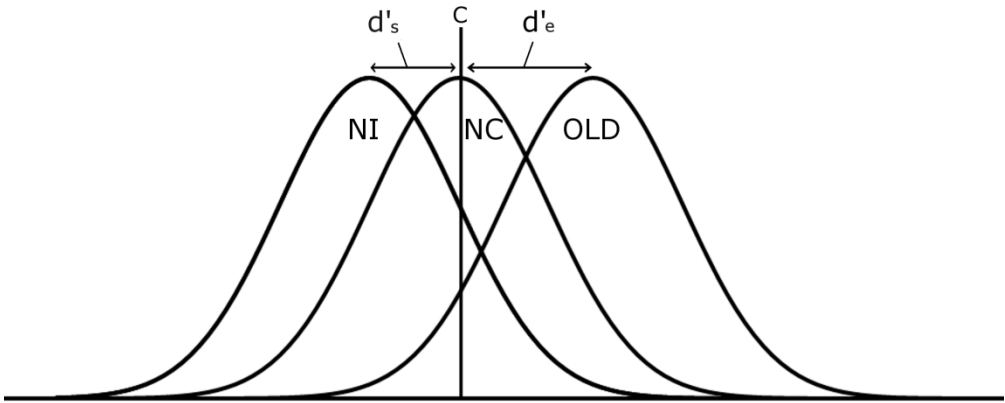


Figure 1 – Signal detection model for Experiments 1 to 3. Participants are assumed to rely on a yes/no criterion (C) applied to the strength-of-evidence dimension to make both recognition and classification judgments. Old = words seen at study, new = words not seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words, d'_s = Structural signal detection measure d' , d'_e = episodic signal detection measure d' .

194x95mm (300 x 300 DPI)

Table 1

Mean (SE) Endorsement Rates by Strength, Task Type and Word Type in Experiment 1

| Task and word type | Strength | |
|--------------------|-----------|-----------|
| | Weak | Strong |
| Recognition | | |
| Old | .88 (.02) | .92 (.02) |
| NC | .18 (.03) | .15 (.02) |
| NI | .13 (.03) | .11 (.02) |
| Classification | | |
| Old | .87 (.02) | .93 (.02) |
| NC | .32 (.04) | .31 (.04) |
| NI | .26 (.04) | .22 (.03) |

Note: Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 2
Mean Episodic Effect (d'_e) and Structural Effect (d'_s) as a Function of Study Strength and Task Type in Experiment 1 (SE in brackets)

| Task and d' type | Strength | | |
|--------------------|-------------|--------------|-------------|
| | Weak | Strong | Total |
| Recognition | | | |
| d'_e | 2.36* (.14) | 2.80* (.18) | 2.58* (.11) |
| d'_s | .30* (.10) | .24* (.06) | .27* (.06) |
| Classification | | | |
| d'_e | 1.88* (.18) | 2.29* (.11)* | 2.09* (.11) |
| d'_s | .21* (.09) | .30* (.09)* | .26* (.06) |
| Both tasks | | | |
| d'_e | 2.12* (.13) | 2.55* (.13) | 2.33* (.09) |
| d'_s | .25* (.07) | .27* (.07) | .26* (.05) |

Note: * = Lower bound of 95% confidence interval above chance level of 0.

Table 3

Final Questionnaire Section Mean Endorsement Rates for Experiments 1, 2, 3 and 4

| Criterion | Experiment 1 Endorsements | Experiment 2 Endorsements | Experiment 3 Endorsements | Experiment 4 Endorsements |
|-------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| Length | 0.10 | 0.14 | 0.13 | 0.21 |
| Syllables | 0.07 | 0.05 | 0.08 | 0.07 |
| Category (noun/verb) | 0.24 | 0.44 | 0.23 | 0.34 |
| Letters | 0.02 | 0.11 | 0.05 | 0.07 |
| Meaning | 0.31 | 0.25 | 0.34 | 0.30 |
| Familiarity | 0.38 | 0.11 | 0.17 | 0.15 |
| Frequency | 0.38 | 0.25 | 0.34 | 0.20 |
| Association | 0.36 | 0.39 | 0.48 | 0.46 |
| Position | 0.07 | 0.18 | 0.06 | 0.09 |
| Concreteness | 0.07 | 0.11 | 0.11 | 0.10 |

Note: Endorsement rates are calculated as a proportion of the total possible selections for that criterion. Because participants were required to select 2 criteria, these rates sum to 2.

Table 4

Mean Chan’s Difference Scores Comparing NC with NI Endorsement Confidence by Strength, and Task Type in Experiment 1 (SE in brackets)

| Task | Weak | Strong | Total |
|----------------|--------------|--------------|--------------|
| Recognition | 4.17* (1.93) | 4.38* (2.05) | 4.28* (1.41) |
| Classification | 0.31 (1.62) | 4.35* (1.72) | 2.33* (1.18) |
| Both Tasks | 2.24 (1.39) | 4.37* (1.48) | 3.30* (1.02) |

Note: * = Lower bound of 95% confidence interval above chance level of 0. NC = new consistent; NI = new inconsistent.

Table 5

Mean Structural Effect by Task Type, Study Strength and Attribution in Experiment 1 (SE in brackets)

| Task Type | Weak study | | | | Strong study | | | |
|----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|-----------------|----------------|
| | Random Chance | Intuition | Memory | Rules | Random Chance | Intuition | Memory | Rules |
| Recognition | .008 (.009) | .015 (.013) | .041* (.010)* | -.004 (.004) | -.003 (.009) | .030* (.013) | .013 (.010) | .001 (.004) |
| Classification | -.010 (.012) | .042* (.017) | .019 (.011) | .015 (.014) | .010 (.012) | .028 (.017) | .036* (.011) | .016 (.014) |

Note: * = Lower bound of 95% confidence interval above chance level of 0.

Table 6
Mean (SE) Endorsement Rates by Strength, Word Type and Task Type from Experiment 2

| Task and word type | Strength | |
|--------------------|-----------|-----------|
| | Weak | Strong |
| Recognition | | |
| Old | .71 (.04) | .80 (.02) |
| NC | .34 (.04) | .26 (.02) |
| NI | .26 (.04) | .18 (.03) |
| Classification | | |
| Old | .67 (.04) | .80 (.04) |
| NC | .40 (.05) | .41 (.04) |
| NI | .33 (.05) | .36 (.04) |

Note: Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 7

Mean Episodic Effect (d'_e) and Structural Effect (d'_s) as a Function of Study Strength and Task Type in Experiment 2 (SE in brackets)

| Task and d' type | Strength | | |
|--------------------|-------------|-------------|-------------|
| | Weak | Strong | Total |
| Recognition | | | |
| d'_e | 1.04* (.10) | 1.54* (.11) | 1.39* (.08) |
| d'_s | .29* (.07) | .35* (.09) | .32* (.06) |
| Classification | | | |
| d'_e | .76* (.10) | 1.20* (.15) | .98* (.08) |
| d'_s | .17 (.11) | .14 (.08) | .16* (.06) |
| Both Tasks | | | |
| d'_e | .91* (.07) | 1.38* (.09) | 1.14* (.06) |
| d'_s | .23* (.06) | .25* (.06) | .24* (.04) |

Note: * = Lower bound of 95% confidence interval above chance level of 0.

Table 8

Mean Chan’s Difference Scores Comparing NC with NI Endorsement Confidence by Strength, and Task Type in Experiment 2 (SE in brackets)

| Task | Weak | Strong | Total |
|----------------|--------------|--------------|-------------|
| Recognition | -2.09 (1.74) | 4.48* (1.74) | 1.20 (1.23) |
| Classification | -0.48 (1.87) | 0.86 (1.87) | 0.19 (1.32) |
| Both Tasks | -1.28 (1.28) | 2.67* (1.28) | 0.69 (0.90) |

Note: * = Lower bound of 95% confidence interval above chance level of 0. NC = new consistent; NI = new inconsistent.

Table 9

Mean Structural Effect by Task Type, Study Strength and Attribution in Experiment 2 (SE in brackets)

| Task Type | Weak study | | | | Strong study | | | |
|----------------|------------------|-----------------|-----------------|------------------|------------------|----------------|-----------------|-----------------|
| | Random Chance | Intuition | Memory | Rules | Random Chance | Intuition | Memory | Rules |
| Recognition | .022 (.015) | .047* (.018) | .013 (.018) | .005 (.013) | .014 (.015) | .029 (.018) | .033 (.018) | .015 (.014) |
| Classification | .023 (.016) | .029 (.018) | -.012 (.020) | 0.053* (.014) | .004 (.016) | .005 (.020) | -.004 (.020) | .035* (.014) |

Note: * = Lower bound of 95% confidence interval above chance level of 0.

Table 10
Mean (SE) Endorsement Rates by Strength, Word Type and Task Type in Experiment 3

| Task and word type | Strength | |
|--------------------|-----------|-----------|
| | Weak | Strong |
| Recognition | | |
| Old | .67 (.03) | .81 (.03) |
| NC | .34 (.03) | .22 (.03) |
| NI | .24 (.03) | .15 (.03) |
| Classification | | |
| Old | .70 (.03) | .86 (.04) |
| NC | .50 (.05) | .36 (.05) |
| NI | .38 (.05) | .25 (.04) |

Note: Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 11

The Episodic Effect (d'_e) and Structural Effect (d'_s) as a Function of Strength and Task Type in Experiment 3 (SE in brackets)

| Task and d' type | Strength | | |
|--------------------|------------|-------------|-------------|
| | Weak | Strong | Total |
| Recognition | | | |
| d'_e | .90* (.09) | 1.89* (.22) | 1.38* (.15) |
| d'_s | .32* (.10) | .29* (.10) | .30* (.07) |
| Classification | | | |
| d'_e | .58* (.13) | 1.80* (.23) | 1.19* (.17) |
| d'_s | .34* (.08) | .33* (.10) | .33* (.06) |
| Both Tasks | | | |
| d'_e | .74* (.08) | 1.85* (.16) | 1.29* (.11) |
| d'_s | .33* (.06) | .31* (.07) | .32* (.05) |
| Total | .53* (.07) | 1.08* (.07) | .81* (.04) |

Note: * = Lower bound of 95% confidence interval above chance level of 0.

Table 12

Mean Chan’s Difference Scores Comparing NC with NI Endorsement Confidence by Strength, and Task Type in Experiment 3 (SE in brackets)

| Task | Weak | Strong | Total |
|----------------|-------------|--------------|--------------|
| Recognition | 0.64 (1.94) | 6.51* (2.01) | 3.58* (1.40) |
| Classification | 0.88 (1.94) | 4.49* (1.94) | 2.69* (1.37) |
| Both Tasks | 0.76 (1.37) | 5.50* (1.40) | 3.13* (0.98) |

Note: * = Lower bound of 95% confidence interval above chance level of 0. NC = new consistent; NI = new inconsistent.

Table 13

Mean Structural Effect by Task Type, Study Strength and Attribution in Experiment 3 (SE in brackets)

| Task Type | Weak study | | | | Strong study | | | |
|----------------|------------------|-----------------|-----------------|-----------------|------------------|----------------|-----------------|-----------------|
| | Random Chance | Intuition | Memory | Rules | Random Chance | Intuition | Memory | Rules |
| Recognition | .019 (.013) | .061* (.018) | .003 (.015) | .005 (.014) | .009 (.014) | .011 (.018) | .041* (.015) | .003 (.014) |
| Classification | .000 (.013) | .031 (.018) | .039* (.015) | .050* (.014) | .011 (.013) | .031 (.018) | .016 (.015) | .049* (.014) |

Note: * = Lower bound of 95% confidence interval above chance level of 0.

Table 14
Mean (SE) Endorsement Rates by Strength, Word Type and Variety in Experiment 4

| Task and Variety | Strength | |
|------------------|-----------|-----------|
| | Weak | Strong |
| 10 Words | | |
| NC | .39 (.02) | .34 (.02) |
| NI | .35 (.02) | .29 (.02) |
| 80 Words | | |
| NC | .40 (.02) | .46 (.03) |
| NI | .34 (.02) | .35 (.02) |

Note: NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 15

The Structural Effect (d'_s) as a Function of Strength and Variety in Experiment 4 (SE in brackets)

| Variety type | Strength | | |
|--------------|------------|------------|-------------|
| | Weak | Strong | Total |
| 10 words | .12* (.05) | .14* (.05) | 0.13* (.04) |
| 80 words | .18* (.06) | .30* (.06) | 0.24* (.04) |
| Total | .15* (.04) | .22* (.04) | |

Note: * = Lower bound of 95% confidence interval above chance level of 0. NC = new consistent; NI = new inconsistent.

Table 16

Mean Chan’s Difference Scores Comparing NC with NI Endorsement Confidence by Strength, and Variety in Experiment 4 (SE in brackets)

| Task | Weak | Strong | Total |
|----------|--------------|--------------|--------------|
| 10 Words | -0.25 (0.90) | 2.81* (0.90) | 1.28* (0.63) |
| 80 Words | -0.17 (0.97) | 4.49* (1.03) | 1.45* (0.71) |
| Total | -0.21 (0.66) | 2.94* (0.68) | |

Note: * = Lower bound of 95% confidence interval above chance level of 0.

Table 17

Mean Structural Effect by Task Type, Study Strength and Attribution in Experiment 4 (SE in brackets)

| Task Type | Weak study | | | | Strong study | | | |
|-----------|------------------|----------------|-----------------|-----------------|------------------|----------------|-----------------|-----------------|
| | Random Chance | Intuition | Memory | Rules | Random Chance | Intuition | Memory | Rules |
| 10 Words | -0.02* (.009) | 0.02 (.011) | 0.02* (.009) | 0.03* (.011) | 0.00 (.006) | 0.02 (.011) | 0.01 (.010) | 0.02 (0.013) |
| 80 Words | 0.02* (.008) | .010 (.015) | 0.02 (.012) | 0.01 (.015) | 0.01 (.009) | 0.01 (.015) | 0.05* (.015) | 0.05* (.017) |

Note: * = Lower bound of 95% confidence interval above chance level of 0.